KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

Kateřina Kratochvílová

Diplomová práce

Inženýrská informatika
Medicínská informatika
2019/2020

Vedoucí práce:
Ing. Lucie Houdová Ph.D.

Nástroj pro automatickou identifikaci KIR alel

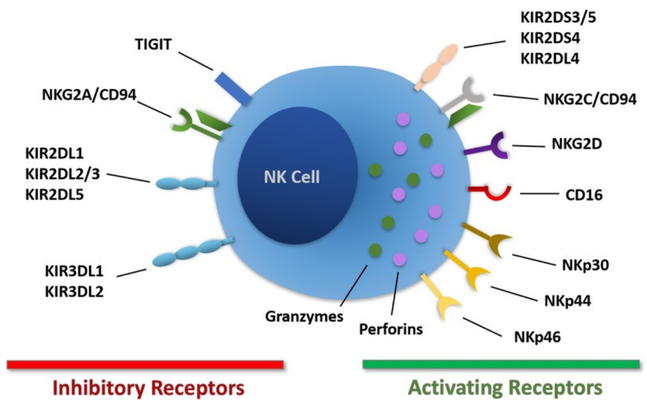
Abstrakt

Diplomová práce se zabývá identifikací KIR alel. Cílem práce je návrh a implementace nástroje pro jejich automatickou identifikaci. V práci jsou představeny KIR geny a metody získávání genomických dat s využitím DNA sekvenace, konkrétně next-generation sequencing (NGS). Dále byly analyzovány využitelné bioinformatické nástroje. Samotný identifikační nástroj byl vyvíjen na readech a nakonec testován a verifikován na datech komerčních linii DNA získaných z FN Plzeň/ BC LF UK Plzeň. Vytváření syntetických readů probíhalo pomocí nástroje ART, Pro zarovnání readů na referenční DNA sekvence byl využit nástroj Bowtie2. V rámci vývoje bylo navrženo několik možných přístupů, které byly poté vyhodnoceny s ohledem na jejich možné využití

Úvod

Transplantace krvetvorných buněk je proces při kterém jsou dárci odebrány krvetvorné buňky, které jsou následně vpraveny do těla pacienta trpícím hematologickou poruchou (například akutní myeloidní leukemie). K potlačení po-transplantačních komplikací se vybírají dárci podle HLA znaků. Prokázan i vliv KIR (Killer-cell immunoglobulin-like receptor) genů, který snižuje riziko návratu nemoci. Aktuálním tématem je možnost vlivu KIR alel. (Alela je konkrétní forma genu). K určení HLA a KIR znaků se využívají sekvenační metody.

Natural killer je buňka imunitního systému a pomáhá mu identifikovat a odstraňovat buňky infikované virem či buňky transformované v nádorové. Produkty vzniklé z KIR genů se nacházejí právě na této buňce.



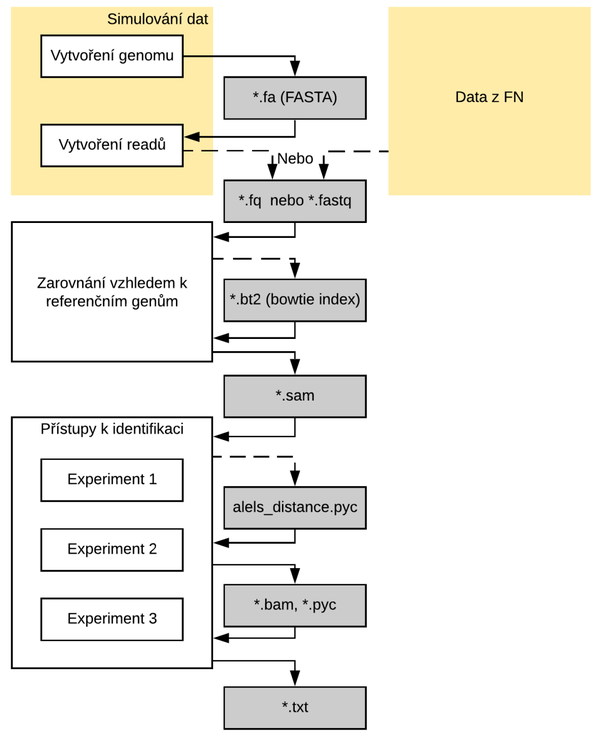
Natural killer buňka a její receptory.

Východiska, analytická část

Výstupem ze sekvenátoru a tedy i hlavní vstup nástroje, jsou ready, které je možné si představit jako posloupnost písmen A, C, G a T dlouhých 250 bp (zjednodušeně 250 znaků). Tyto ready je třeba zarovnat vůči referenční sekvenci. Jedna alela může být dlouhá téměř 16000 bp. O délce readů rozhoduje zvolená sekvenační metoda.V tomto případě Illumina. Ready musejí být zarovnány na referenční sekvenci. Následně je třeba vyhodnotit, která alela se v genomu nacházela či nikoliv.

Hlavní aspekty realizace

Návrh pipeliny:



Vyhodnocení zarovnání:

1. Odstranění alel, které mají menší šířku pokrytí. (zarovnalo se na ně málo readů).
2. U blízkých alel je třeba rozhodnout která z nich to může být. Porovnání hloubky pokrytí v místech kde se alely neshodují.
3. U některých genů se u alel, které do genomu patří objevili vysoké vrcholy. Identifikace této alely a odstranění alel bez tohoto vrcholu patřící do stejného genu.

Problémy:

- Ne všechny alely mají známou refrenční sekvenci (v současnosti známo 1109 alel, referenční sekvence dostupná pro 461 alel).
- Podobnost alel: nezřídka je read zarovnán na jinou alelu než ke které patří. Nejmenší vzdálenost mezi alelami je 1.
- Při sekvenování vznikají chyby. Může to být změna báze za jinou, báze přečtená dvakrát či chybějící báze.

Dosažené výsledky

V rámci syntetických readů bylo dosahováno až 80 % přesnosti a 87 % úplnosti. V případě reálných dat nebylo možné tyto metriky posoudit vzhledem k tomu, že z biologického hlediska nebylo možné alelu určit až na úroveň non-coding regionu.

Závěr

Nástroj ve většině případů dokáže určit alelu na nějakou úroveň. Dopišu asi až plně zanalyzuju to na jakou úroveň a co to dokáže identifikovat. To samý tady dopsat i s realnym daty?

Vizualizace pokrytí alel. Šířkou pokrytí se rozumí pokrytí po délce, zatímco hloubka pokrytí znační pokrytí do výšky.

FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI